Class 09 Halloween Mini Project

Isabella Ruud: PID A59016138

Table of contents

Today we are delving into an analysis of Halloween Candy data using ggplot, dplyr, basic stats, correlation, and PCA

Section 1: load in the data

Read in the data:

```
candy_file <- "candy-data.txt"

candy = read.csv(candy_file, row.names=1)
head(candy)
```

```
             chocolate fruity caramel peanutyalmondy nougat crispedricewafer
100 Grand            1      0       1              0      0                 1
3 Musketeers         1      0       0              0      1                 0
One dime             0      0       0              0      0                 0
One quarter          0      0       0              0      0                 0
Air Heads            0      1       0              0      0                 0
Almond Joy           1      0       0              1      0                 0
             hard bar pluribus sugarpercent pricepercent winpercent
100 Grand       0   1        0        0.732        0.860   66.97173
```

```
3 Musketeers    0   1       0       0.604       0.511   67.60294
One dime        0   0       0       0.011       0.116   32.26109
One quarter     0   0       0       0.011       0.511   46.11650
Air Heads       0   0       0       0.906       0.511   52.34146
Almond Joy      0   1       0       0.465       0.767   50.34755
```

Q1. How many different candy types are in this dataset?

```
nrow(candy)
```

```
[1] 85
```

There are 85 different candy types in this dataset

Q2. How many fruity candy types are in the dataset?

```
table(candy$fruity)
```

```
 0  1
47 38
```

```
sum(candy$fruity)
```

```
[1] 38
```

There are 38 fruity candy types in the dataset.

Q2. how many chocolate candy types are in the dataset?

```
sum(candy$chocolate)
```

```
[1] 37
```

There are 37 chocolate candy types in the dataset.

Section 2: what is your favorite candy?

Q3. What is your favorite candy in the dataset and what is it's winpercent value?

```
candy["Reese's pieces",]$winpercent
```

```
[1] 73.43499
```

```
candy["Reese's pieces","winpercent"]
```

```
[1] 73.43499
```

After looking at the list of candy in the dataset, Twix is my favorite candy in the dataset and it has a winpercent value of 81.64291

We can also use the filter() and select() functions from dplyr

```
library(dplyr)
candy |>
  filter(rownames(candy) == "Twix") |>
  select(winpercent)
```

```
     winpercent
Twix   81.64291
```

```
candy |>
  filter(rownames(candy) == "Twix") |>
  select(winpercent,sugarpercent)
```

```
     winpercent sugarpercent
Twix   81.64291        0.546
```

```
candy |>
  filter(rownames(candy) == "Nerds") |>
  select(winpercent,sugarpercent)
```

```
      winpercent sugarpercent
Nerds   55.35405        0.848
```

Q4. What is the winpercent value for "Kit Kat"?

```
candy['Kit Kat',]$winpercent
```

```
[1] 76.7686
```

The winpercent for Kit Kat is 76.7686

Q5. What is the winpercent value for “Tootsie Roll Snack Bars”?

```
candy['Tootsie Roll Snack Bars',]$winpercent
```

```
[1] 49.6535
```

The winpercent for Tootsie Roll Snack Bars is 49.6535

A useful function for a quick look at a new dataset is found in the skimr package.

```
#library("skimr")
skimr::skim(candy)
```

Table 1: Data summary

Name	candy
Number of rows	85
Number of columns	12
_____	
Column type frequency:	
numeric	12
_____	
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
chocolate	0	1	0.44	0.50	0.00	0.00	0.00	1.00	1.00	
fruity	0	1	0.45	0.50	0.00	0.00	0.00	1.00	1.00	
caramel	0	1	0.16	0.37	0.00	0.00	0.00	0.00	1.00	
peanutyalmondy	0	1	0.16	0.37	0.00	0.00	0.00	0.00	1.00	
nougat	0	1	0.08	0.28	0.00	0.00	0.00	0.00	1.00	
crispedricewafer	0	1	0.08	0.28	0.00	0.00	0.00	0.00	1.00	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
hard	0	1	0.18	0.38	0.00	0.00	0.00	0.00	1.00	
bar	0	1	0.25	0.43	0.00	0.00	0.00	0.00	1.00	
pluribus	0	1	0.52	0.50	0.00	0.00	1.00	1.00	1.00	
sugarpercent	0	1	0.48	0.28	0.01	0.22	0.47	0.73	0.99	
pricepercent	0	1	0.47	0.29	0.01	0.26	0.47	0.65	0.98	
winpercent	0	1	50.32	14.71	22.45	39.14	47.83	59.86	84.18	

Q6. Is there any variable/column that looks to be on a different scale to the majority of the other columns in the dataset?

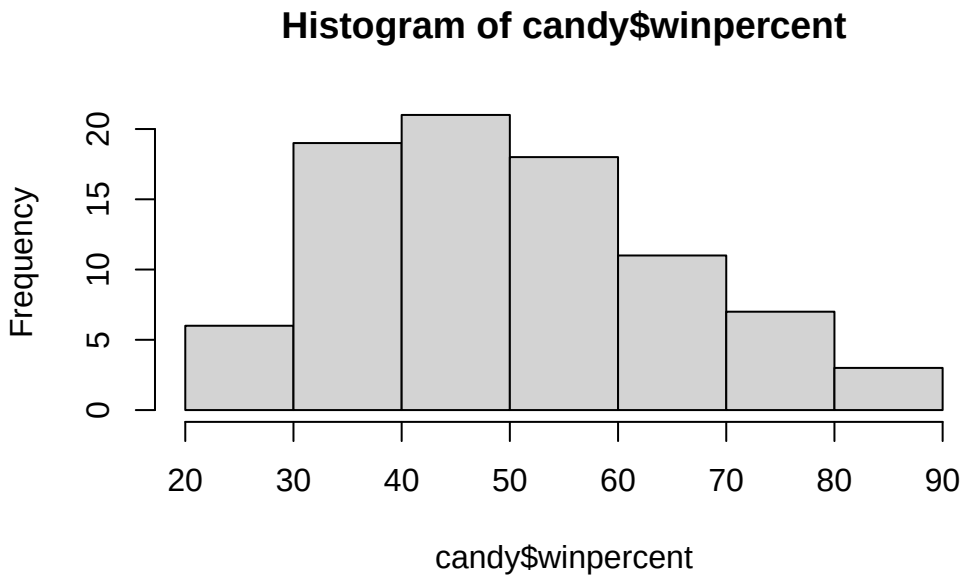
The winpercent variable seems to be a different scale than the majority of the other columns since it ranges from 0 to 100. Most columns are either 0 or 1 values. The pricepercent and sugarpercent columns have values ranging from 0 to 1. Because of this, we should scale the data before analysis like PCA.

Q7. What do you think a zero and one represent for the candy$chocolate column?

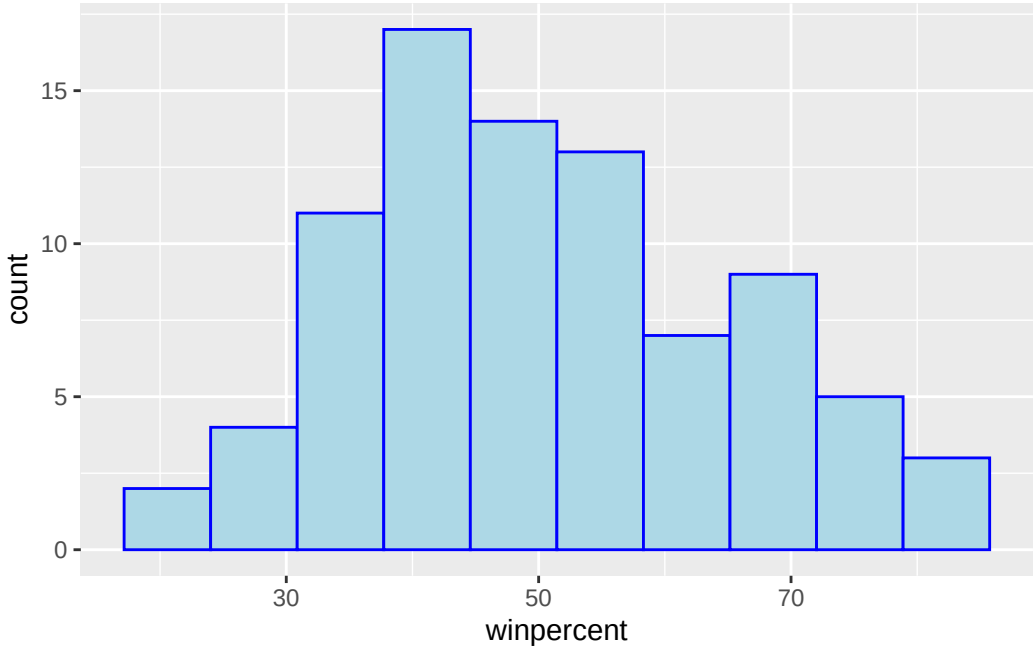
They represent true (1) and false (0) values, so 1 means that the candy contains chocolate and 0 means that the candy does not contain chocolate.

Q8. Plot a histogram of winpercent values use base R and ggplot

```
hist(candy$winpercent)
```



```
library(ggplot2)
ggplot(candy) + aes(x=winpercent) + geom_histogram(bins = 10, fill = "lightblue", color = "bl
```



Q9. Is the distribution of winpercent values symmetrical?

No, the distribution is not symmetrical. The distribution of winpercent values is slightly skewed to the right since it has a center towards the lower winpercent values and it has a tail at the higher winpercent values

Q10. Is the center of the distribution above or below 50%?

The center of the distribution is below 50%

```
summary(candy$winpercent)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 22.45   39.14   47.83   50.32   59.86   84.18
```

Q11. On average is chocolate candy higher or lower ranked than fruit candy?

Step 1. find extract chocolate candy rows in the dataset Step 2. get their winpercent values step 3. caculate their mean winpercent step4/5/6. repeat for fruity candy step 7. compare mean chocolate winpercent to mean fruity win percent

```r
#step 1
choc.inds <- candy$chocolate == 1
choc.candy <- candy[choc.inds,]

#step 2
choc.win <- choc.candy$winpercent

#step 3
choc.mean <- mean(choc.win)

#steps4/5/6
fruity.inds <- candy$fruity == 1
fruity.candy <- candy[fruity.inds,]
fruity.win <- fruity.candy$winpercent
fruity.mean <- mean(fruity.win)

paste("chocolte: ", choc.mean)
```

```
[1] "chocolte:  60.9215294054054"
```

```r
paste("fruity: ", fruity.mean)
```

```
[1] "fruity:  44.1197414210526"
```

```r
chocolate_win_mean <- mean(candy$winpercent[as.logical(candy$chocolate)])
fruity_win_mean <- mean(candy$winpercent[as.logical(candy$fruity)])

paste("Chocolate:", chocolate_win_mean)
```

```
[1] "Chocolate: 60.9215294054054"
```

```r
paste("Fruity:", fruity_win_mean)
```

```
[1] "Fruity: 44.1197414210526"
```

On average, chocolate candy has a higher win percentage than fruity candy

Q12. Is this difference statistically significant?

Let's use

```
chocolate_win <- candy$winpercent[as.logical(candy$chocolate)]
fruity_win <- candy$winpercent[as.logical(candy$fruity)]

t.test(choc.win, fruity.win)
```

```
    Welch Two Sample t-test

data:  choc.win and fruity.win
t = 6.2582, df = 68.882, p-value = 2.871e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 11.44563 22.15795
sample estimates:
mean of x mean of y
 60.92153  44.11974
```

The difference between the win percentage for chocolate candy and fruity candy is significant since the p value is less than 0.05 (p-value = 2.871e-08)

Section 3: overall candy rankings

Q13. What are the five least liked candy types in this set?

I can use the output of order(winpercent) to re-arrange (or order) my whole dataset by winpercent

```
ord.inds <- order(candy$winpercent)
head(candy[ord.inds,], n = 6)
```

```
                  chocolate fruity caramel peanutyalmondy nougat
Nik L Nip                 0      1       0              0      0
Boston Baked Beans        0      0       0              1      0
Chiclets                  0      1       0              0      0
Super Bubble              0      1       0              0      0
Jawbusters                0      1       0              0      0
Root Beer Barrels         0      0       0              0      0
                  crispedricewafer hard bar pluribus sugarpercent pricepercent
Nik L Nip                        0    0   0        1        0.197        0.976
Boston Baked Beans               0    0   0        1        0.313        0.511
```

```
Chiclets                        0    0   0        1        0.046        0.325
Super Bubble                    0    0   0        0        0.162        0.116
Jawbusters                      0    1   0        1        0.093        0.511
Root Beer Barrels               0    1   0        1        0.732        0.069
                   winpercent
Nik L Nip            22.44534
Boston Baked Beans   23.41782
Chiclets             24.52499
Super Bubble         27.30386
Jawbusters           28.12744
Root Beer Barrels    29.70369
```

```
candy |>
  arrange(winpercent) |>
  head(5)
```

```
                   chocolate fruity caramel peanutyalmondy nougat
Nik L Nip                  0      1       0              0      0
Boston Baked Beans         0      0       0              1      0
Chiclets                   0      1       0              0      0
Super Bubble               0      1       0              0      0
Jawbusters                 0      1       0              0      0
                   crispedricewafer hard bar pluribus sugarpercent pricepercent
Nik L Nip                         0    0   0        1        0.197        0.976
Boston Baked Beans                0    0   0        1        0.313        0.511
Chiclets                          0    0   0        1        0.046        0.325
Super Bubble                      0    0   0        0        0.162        0.116
Jawbusters                        0    1   0        1        0.093        0.511
                   winpercent
Nik L Nip            22.44534
Boston Baked Beans   23.41782
Chiclets             24.52499
Super Bubble         27.30386
Jawbusters           28.12744
```

The 5 least liked candies in the dataset are: Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, and Jawbusters (in order from least liked to most liked)

I like using the dplyr version better since it is a little more intuitive

Q14. What are the top 5 all time favorite candy types out of this set?

```
candy |>
  arrange(-winpercent) |>
  head(5)
```

```
                         chocolate fruity caramel peanutyalmondy nougat
Reese's Peanut Butter cup         1      0       0              1      0
Reese's Miniatures                1      0       0              1      0
Twix                              1      0       1              0      0
Kit Kat                           1      0       0              0      0
Snickers                          1      0       1              1      1
                         crispedricewafer hard bar pluribus sugarpercent
Reese's Peanut Butter cup                0    0   0        0        0.720
Reese's Miniatures                       0    0   0        0        0.034
Twix                                     1    0   1        0        0.546
Kit Kat                                  1    0   1        0        0.313
Snickers                                 0    0   1        0        0.546
                         pricepercent winpercent
Reese's Peanut Butter cup        0.651   84.18029
Reese's Miniatures               0.279   81.86626
Twix                             0.906   81.64291
Kit Kat                          0.511   76.76860
Snickers                         0.651   76.67378
```

```
head(candy[order(candy$winpercent, decreasing = TRUE),], n = 5)
```

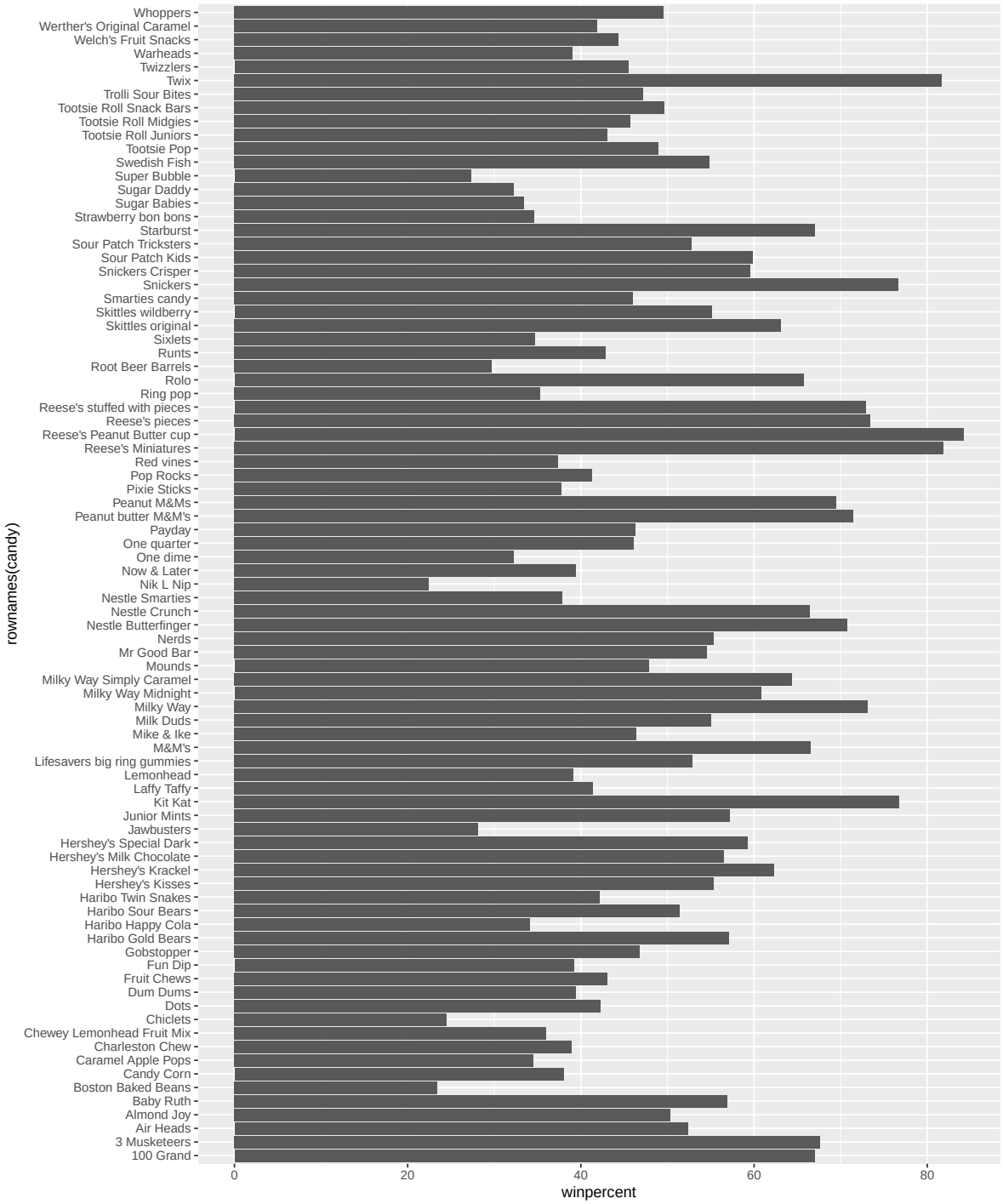
```
                         chocolate fruity caramel peanutyalmondy nougat
Reese's Peanut Butter cup         1      0       0              1      0
Reese's Miniatures                1      0       0              1      0
Twix                              1      0       1              0      0
Kit Kat                           1      0       0              0      0
Snickers                          1      0       1              1      1
                         crispedricewafer hard bar pluribus sugarpercent
Reese's Peanut Butter cup                0    0   0        0        0.720
Reese's Miniatures                       0    0   0        0        0.034
Twix                                     1    0   1        0        0.546
Kit Kat                                  1    0   1        0        0.313
Snickers                                 0    0   1        0        0.546
                         pricepercent winpercent
Reese's Peanut Butter cup        0.651   84.18029
Reese's Miniatures               0.279   81.86626
Twix                             0.906   81.64291
```

```
Kit Kat                         0.511   76.76860
Snickers                        0.651   76.67378
```

The top 5 most popular candies are Reese's Peanut Butter cup, Reese's Miniatures, Twix, Kit Kat, and Snickers

Q15. Make a first barplot of candy ranking based on winpercent values.

```
ggplot(candy) +
  aes(x = winpercent, y = rownames(candy)) +
  geom_col()
```

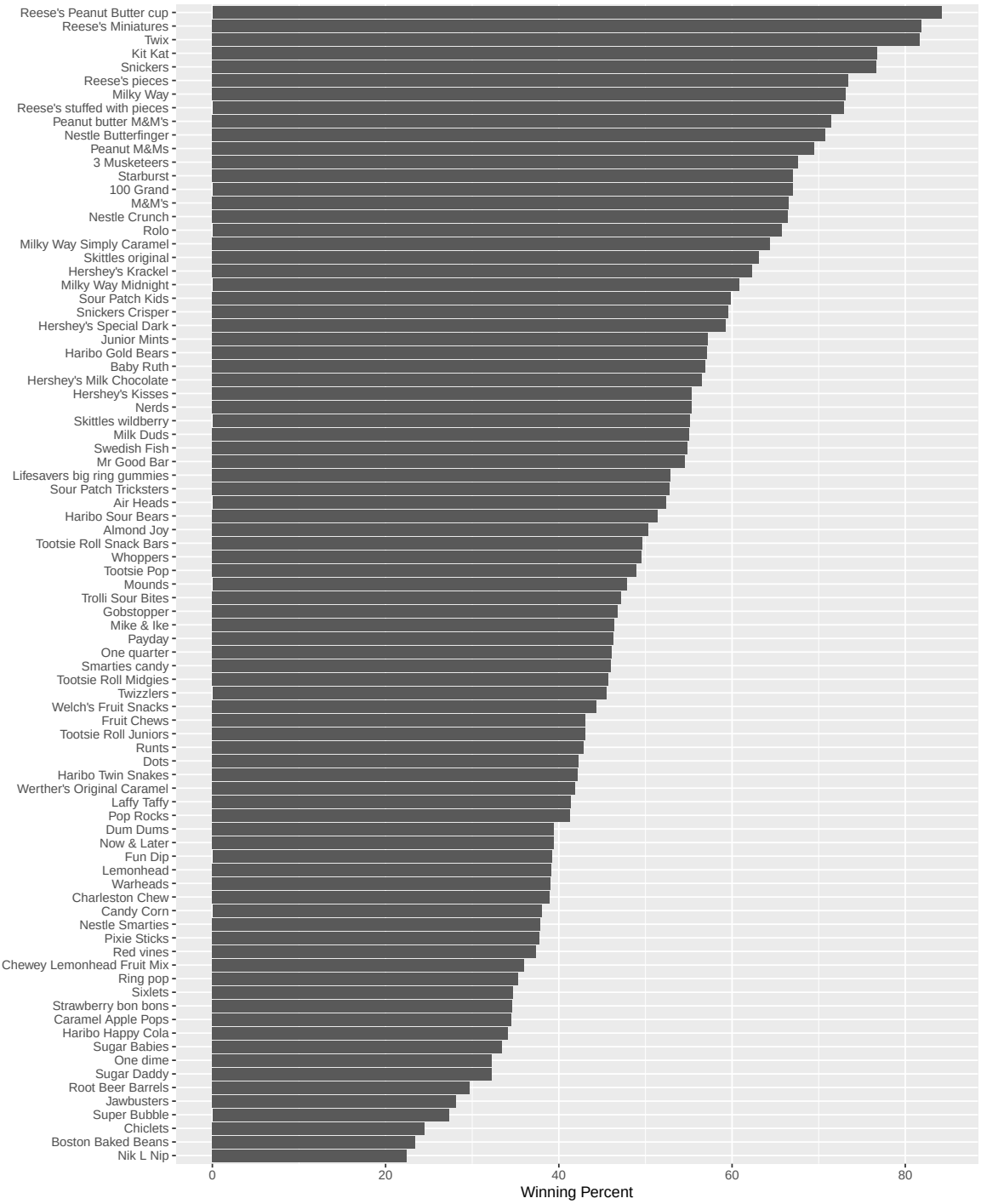


We can make the plot better by rearranging the y axis by winpercent so that the highest scoring candy is at the top and the lowest scoreing candy is at the bottom.

Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

To fix the squished axis issue can change the code chunk itself or save the plot as a png and define the height and weight

```
ggplot(candy) +
  aes(x = winpercent, y = reorder(rownames(candy), winpercent)) +
  geom_col() +
  xlab("Winning Percent") +
  ylab("")
```




```
p <- ggplot(candy) +
  aes(x = winpercent, y = reorder(rownames(candy), winpercent)) +
```

```
  geom_col() +
  xlab("Winning Percent") +
  ylab("")

ggsave("my_plot.png", height=12, width=5)
```

markdown syntax to insert an image:

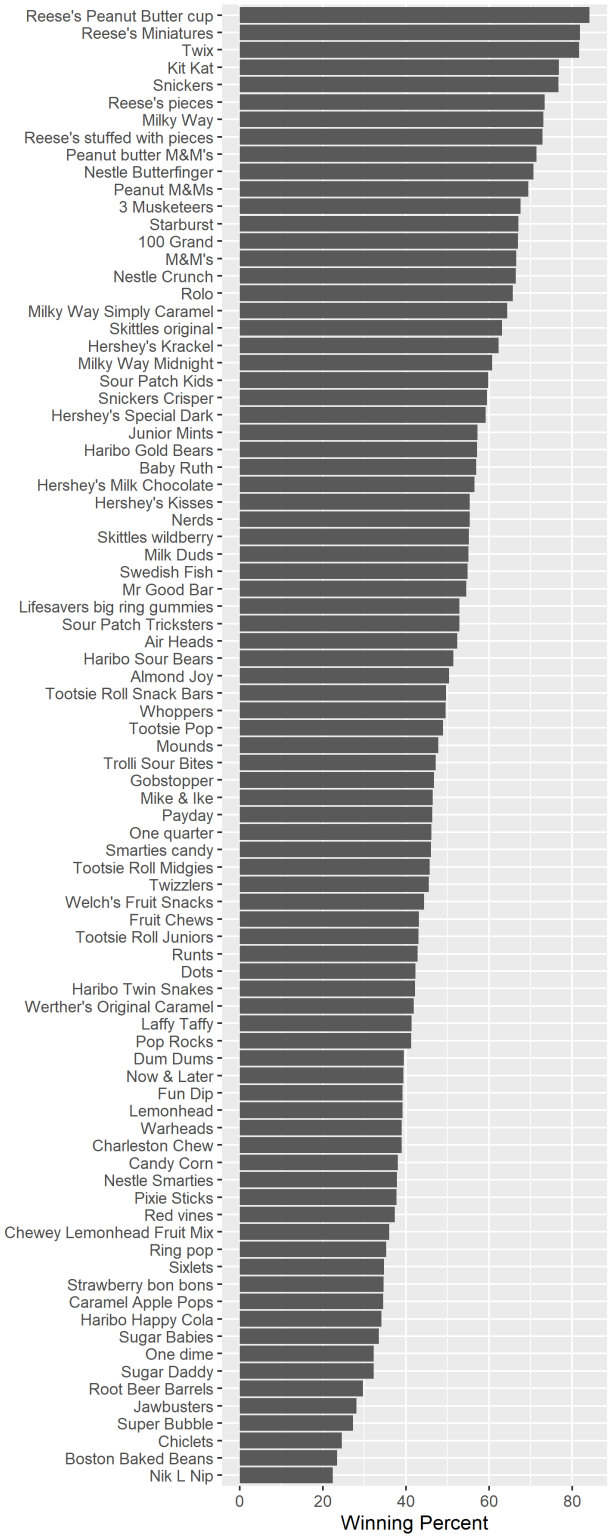
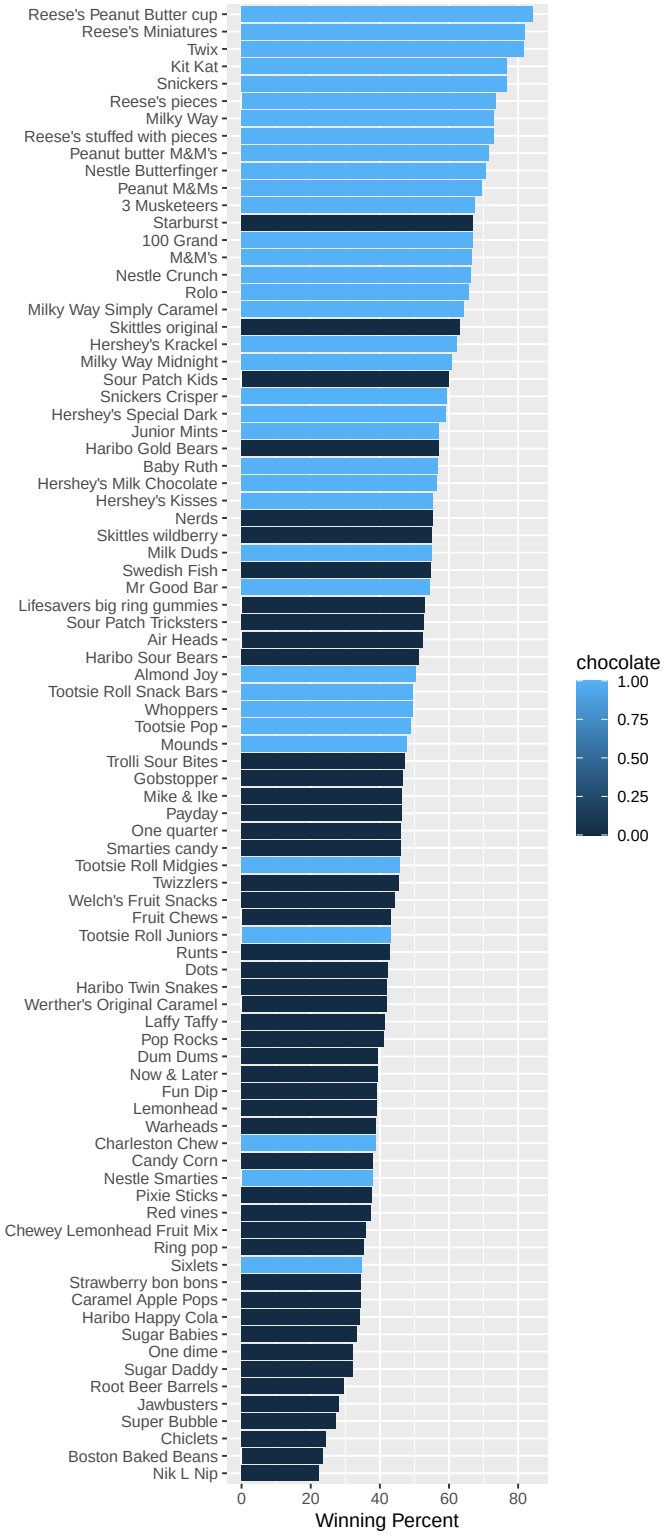


Figure 1: Caption can go here

add some color: color bars by chocolate or not

Can't color by fill=chocolate in aes because it makes it a scale

```
ggplot(candy) +
  aes(x = winpercent, y = reorder(rownames(candy), winpercent), fill = chocolate) +
  geom_col() +
  xlab("Winning Percent") +
  ylab("")
```

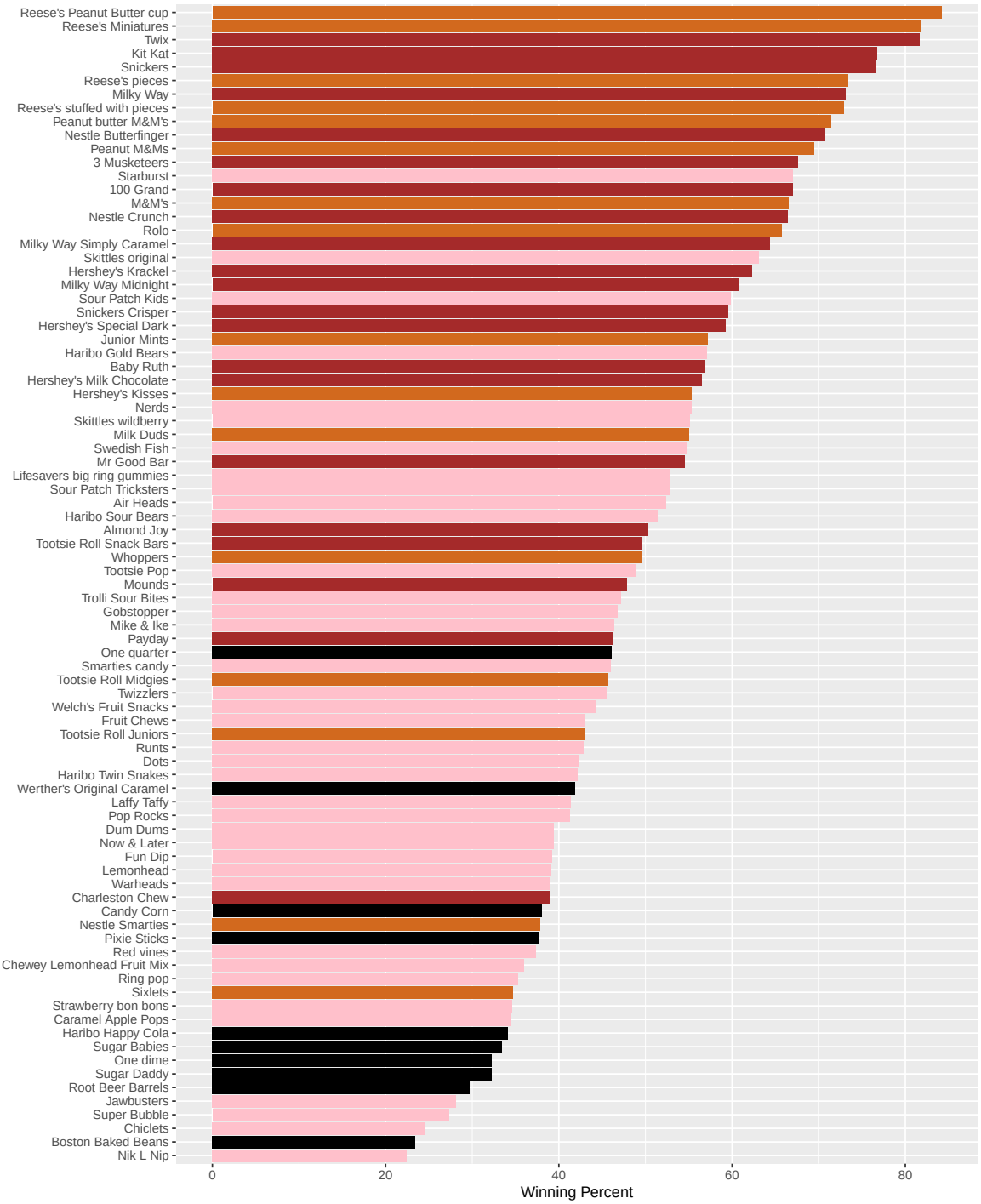

Reese's Peanut Butter cup
Reese's Miniatures
Twix
Kit Kat
Snickers
Reese's pieces
Milky Way
Reese's stuffed with pieces
Peanut butter M&M's
Nestle Butterfinger
Peanut M&Ms
3 Musketeers
Starburst
100 Grand
M&M's
Nestle Crunch
Rolo
Milky Way Simply Caramel
Skittles original
Hershey's Krackel
Milky Way Midnight
Sour Patch Kids
Snickers Crisper
Hershey's Special Dark
Junior Mints
Haribo Gold Bears
Baby Ruth
Hershey's Milk Chocolate
Hershey's Kisses
Nerds
Skittles wildberry
Milk Duds
Swedish Fish
Mr Good Bar
Lifesavers big ring gummies
Sour Patch Tricksters
Air Heads
Haribo Sour Bears
Almond Joy
Tootsie Roll Snack Bars
Whoppers
Tootsie Pop
Mounds
Trolli Sour Bites
Gobstopper
Mike & Ike
Payday
One quarter
Smarties candy
Tootsie Roll Midgies
Twizzlers
Welch's Fruit Snacks
Fruit Chews
Tootsie Roll Juniors
Runts
Dots
Haribo Twin Snakes
Werther's Original Caramel
Laffy Taffy
Pop Rocks
Dum Dums
Now & Later
Fun Dip
Lemonhead
Warheads
Charleston Chew
Candy Corn
Nestle Smarties
Pixie Sticks
Red vines
Chewey Lemonhead Fruit Mix
Ring pop
Sixlets
Strawberry bon bons
Caramel Apple Pops
Haribo Happy Cola
Sugar Babies
One dime
Sugar Daddy
Root Beer Barrels
Jawbusters
Super Bubble
Chiclets
Boston Baked Beans
Nik L Nip
0
20
40
60
80
Winning Percent
chocolate
1.00
0.75
0.50
0.25
0.00

I want to color chocolate and fruity color a specific color

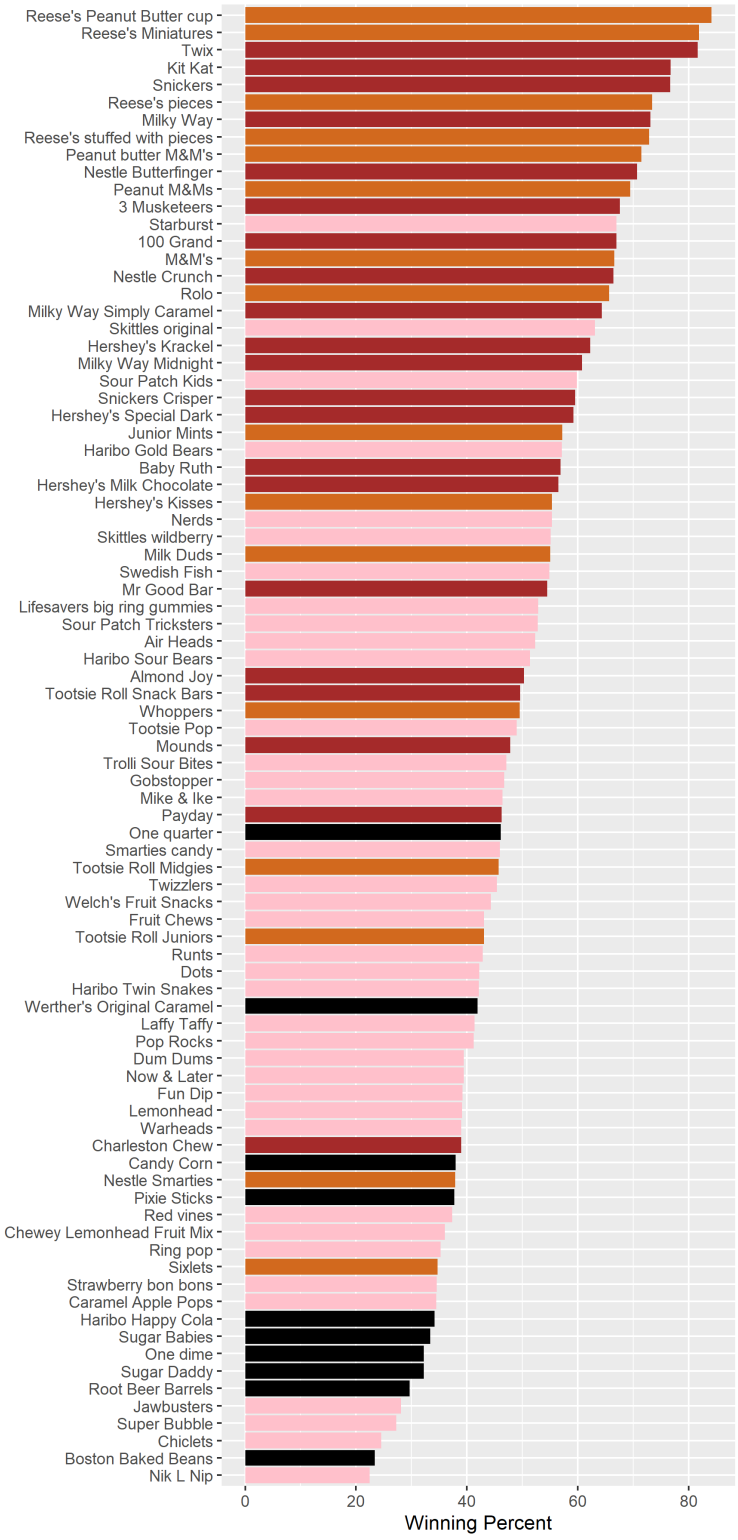
To do this, we need to define our own color map

```
my_cols <- rep("black", nrow(candy))
my_cols[candy$chocolate == 1] <- "chocolate"
my_cols[candy$bar == 1] <- "brown"
my_cols[candy$fruity == 1] <- "pink"


ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +
  geom_col(fill=my_cols) +
  xlab("Winning Percent") +
  ylab("")
```


Reese's Peanut Butter cup
Reese's Miniatures
Twix
Kit Kat
Snickers
Reese's pieces
Milky Way
Reese's stuffed with pieces
Peanut butter M&M's
Nestle Butterfinger
Peanut M&Ms
3 Musketeers
Starburst
100 Grand
M&M's
Nestle Crunch
Rolo
Milky Way Simply Caramel
Skittles original
Hershey's Krackel
Milky Way Midnight
Sour Patch Kids
Snickers Crisper
Hershey's Special Dark
Junior Mints
Haribo Gold Bears
Baby Ruth
Hershey's Milk Chocolate
Hershey's Kisses
Nerds
Skittles wildberry
Milk Duds
Swedish Fish
Mr Good Bar
Lifesavers big ring gummies
Sour Patch Tricksters
Air Heads
Haribo Sour Bears
Almond Joy
Tootsie Roll Snack Bars
Whoppers
Tootsie Pop
Mounds
Trolli Sour Bites
Gobstopper
Mike & Ike
Payday
One quarter
Smarties candy
Tootsie Roll Midgies
Twizzlers
Welch's Fruit Snacks
Fruit Chews
Tootsie Roll Juniors
Runts
Dots
Haribo Twin Snakes
Werther's Original Caramel
Laffy Taffy
Pop Rocks
Dum Dums
Now & Later
Fun Dip
Lemonhead
Warheads
Charleston Chew
Candy Corn
Nestle Smarties
Pixie Sticks
Red vines
Chewey Lemonhead Fruit Mix
Ring pop
Sixlets
Strawberry bon bons
Caramel Apple Pops
Haribo Happy Cola
Sugar Babies
One dime
Sugar Daddy
Root Beer Barrels
Jawbusters
Super Bubble
Chiclets
Boston Baked Beans
Nik L Nip
0
20
40
60
80
Winning Percent

```r
ggsave("my_color_plot.png", height = 12, width = 6)
```


Reese's Peanut Butter cup
Reese's Miniatures
Twix
Kit Kat
Snickers
Reese's pieces
Milky Way
Reese's stuffed with pieces
Peanut butter M&M's
Nestle Butterfinger
Peanut M&Ms
3 Musketeers
Starburst
100 Grand
M&M's
Nestle Crunch
Rolo
Milky Way Simply Caramel
Skittles original
Hershey's Krackel
Milky Way Midnight
Sour Patch Kids
Snickers Crisper
Hershey's Special Dark
Junior Mints
Haribo Gold Bears
Baby Ruth
Hershey's Milk Chocolate
Hershey's Kisses
Nerds
Skittles wildberry
Milk Duds
Swedish Fish
Mr Good Bar
Lifesavers big ring gummies
Sour Patch Tricksters
Air Heads
Haribo Sour Bears
Almond Joy
Tootsie Roll Snack Bars
Whoppers
Tootsie Pop
Mounds
Trolli Sour Bites
Gobstopper
Mike & Ike
Payday
One quarter
Smarties candy
Tootsie Roll Midgies
Twizzlers
Welch's Fruit Snacks
Fruit Chews
Tootsie Roll Juniors
Runts
Dots
Haribo Twin Snakes
Werther's Original Caramel
Laffy Taffy
Pop Rocks
Dum Dums
Now & Later
Fun Dip
Lemonhead
Warheads
Charleston Chew
Candy Corn
Nestle Smarties
Pixie Sticks
Red vines
Chewey Lemonhead Fruit Mix
Ring pop
Sixlets
Strawberry bon bons
Caramel Apple Pops
Haribo Happy Cola
Sugar Babies
One dime
Sugar Daddy
Root Beer Barrels
Jawbusters
Super Bubble
Chiclets
Boston Baked Beans
Nik L Nip
0
20
40
60
80
Winning Percent

Q17. What is the worst ranked chocolate candy?

Sixlets are the worst ranked chocolate candy

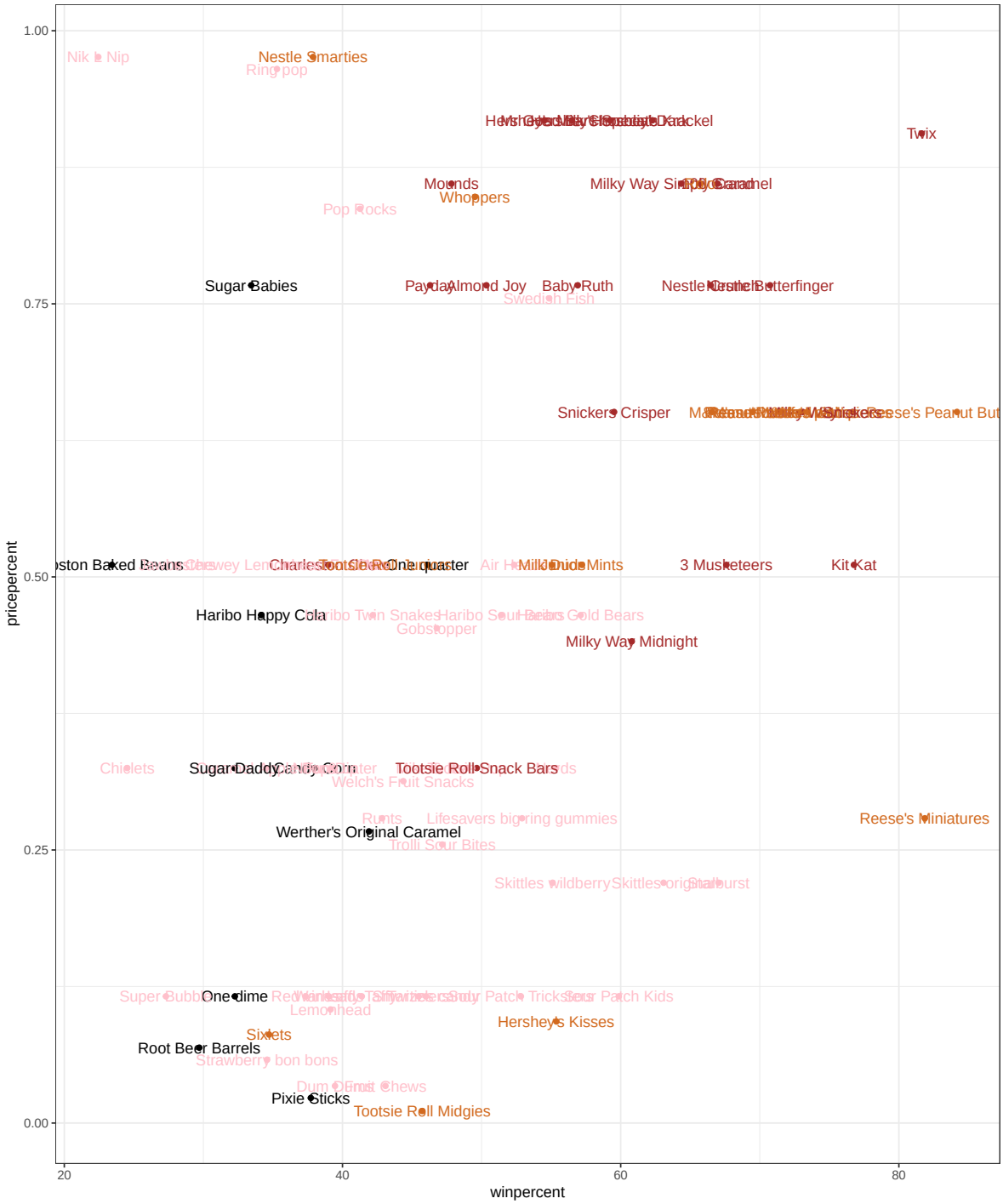
Q18. What is the best ranked fruity candy?

Starburst is the best ranked fruity candy

Section 4: taking a look at pricepoint

Plot of winpercent vs pricepercent

```
ggplot(candy) +
  aes(x= winpercent, y = pricepercent, label = rownames(candy)) +
  geom_point(color = my_cols) +
  theme_bw() +
  geom_text(col = my_cols)
```



Can use ggrepel to avoid label or text overplotting

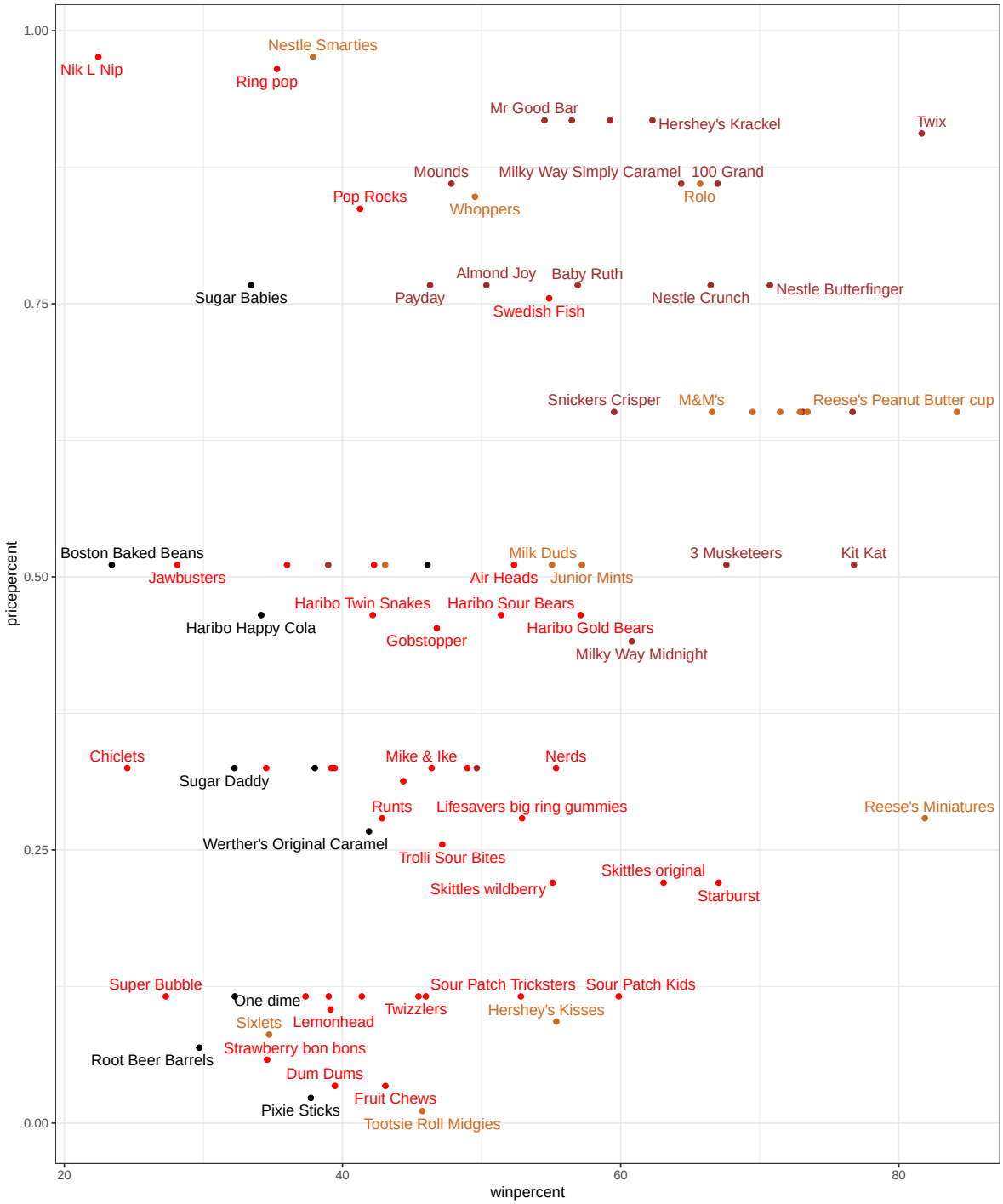
We can control the amount of labels visible setting different max.overlaps values

```
library(ggrepel)

#change the pink to red for fruity candy
my_cols[candy$fruity == 1] = "red"

ggplot(candy) +
  aes(x= winpercent, y = pricepercent, label = rownames(candy)) +
  geom_point(color = my_cols) +
  theme_bw() +
  geom_text_repel(col=my_cols, max.overlaps = 5)
```

```
Warning: ggrepel: 24 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



Q19. Which candy type is the highest ranked in terms of winpercent for the least money - i.e. offers the most bang for your buck?

Reese's miniatures is a good option that has high winpercent for a lower pricepercent

> Q20. What are the top 5 most expensive candy types in the dataset and of these which is the least popular?

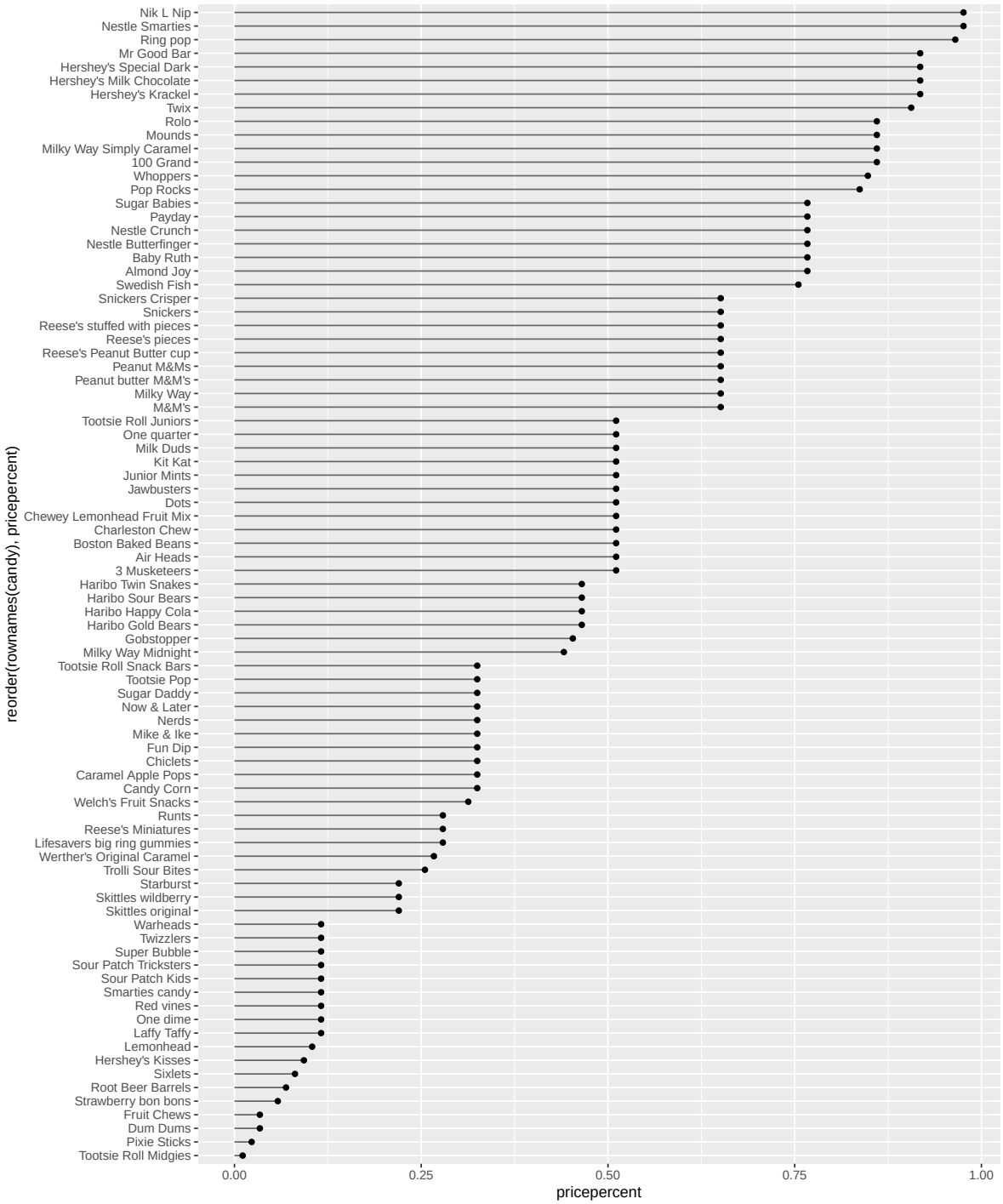
```
head(candy[order(candy$pricepercent, decreasing = TRUE),c("pricepercent", "winpercent")], n=5
```

```
                        pricepercent winpercent
Nik L Nip                      0.976   22.44534
Nestle Smarties                0.976   37.88719
Ring pop                       0.965   35.29076
Hershey's Krackel              0.918   62.28448
Hershey's Milk Chocolate       0.918   56.49050
```

The top five most expensive candies are Nik L Nip, Nestle Smarties, Ring pop, Hershey's Krackel, and Hershey's milk chocolate. Of these, Hershey's Krackel has the highest winpercent so it is the most popular

> Q21. Make a barplot again with geom_col() this time using pricepercent and then improve this step by step, first ordering the x-axis by value and finally making a so called "dot chat" or "lollipop" chart by swapping geom_col() for geom_point() + geom_segment().

```
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
                   xend = 0), col="gray40") +
    geom_point()
```


Nik L Nip
Nestle Smarties
Ring pop
Mr Good Bar
Hershey's Special Dark
Hershey's Milk Chocolate
Hershey's Krackel
Twix
Rolo
Mounds
Milky Way Simply Caramel
100 Grand
Whoppers
Pop Rocks
Sugar Babies
Payday
Nestle Crunch
Nestle Butterfinger
Baby Ruth
Almond Joy
Swedish Fish
Snickers Crisper
Snickers
Reese's stuffed with pieces
Reese's pieces
Reese's Peanut Butter cup
Peanut M&Ms
Peanut butter M&M's
Milky Way
M&M's
Tootsie Roll Juniors
One quarter
Milk Duds
Kit Kat
Junior Mints
Jawbusters
Dots
Chewey Lemonhead Fruit Mix
Charleston Chew
Boston Baked Beans
Air Heads
3 Musketeers
Haribo Twin Snakes
Haribo Sour Bears
Haribo Happy Cola
Haribo Gold Bears
Gobstopper
Milky Way Midnight
Tootsie Roll Snack Bars
Tootsie Pop
Sugar Daddy
Now & Later
Nerds
Mike & Ike
Fun Dip
Chiclets
Caramel Apple Pops
Candy Corn
Welch's Fruit Snacks
Runts
Reese's Miniatures
Lifesavers big ring gummies
Werther's Original Caramel
Trolli Sour Bites
Starburst
Skittles wildberry
Skittles original
Warheads
Twizzlers
Super Bubble
Sour Patch Tricksters
Sour Patch Kids
Smarties candy
Red vines
One dime
Laffy Taffy
Lemonhead
Hershey's Kisses
Sixlets
Root Beer Barrels
Strawberry bon bons
Fruit Chews
Dum Dums
Pixie Sticks
Tootsie Roll Midgies
reorder(rownames(candy), pricepercent)
0.00
0.25
0.50
0.75
1.00
pricepercent

Section 5: exploring the correlation structure

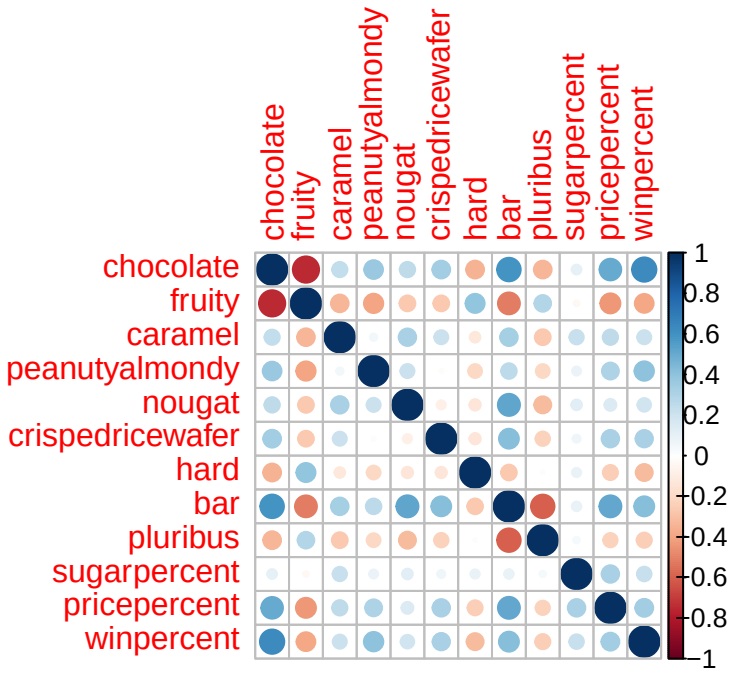
```
library(corrplot)
```

```
corrplot 0.95 loaded
```

```
cij <- cor(candy)
head(cij)
```

```
                  chocolate     fruity     caramel peanutyalmondy      nougat
chocolate         1.0000000 -0.7417211  0.24987535     0.37782357  0.25489183
fruity           -0.7417211  1.0000000 -0.33548538    -0.39928014 -0.26936712
caramel           0.2498753 -0.3354854  1.00000000     0.05935614  0.32849280
peanutyalmondy    0.3778236 -0.3992801  0.05935614     1.00000000  0.21311310
nougat            0.2548918 -0.2693671  0.32849280     0.21311310  1.00000000
crispedricewafer  0.3412098 -0.2693671  0.21311310    -0.01764631 -0.08974359
                  crispedricewafer       hard        bar   pluribus sugarpercent
chocolate               0.34120978 -0.3441769  0.5974211 -0.3396752   0.10416906
fruity                 -0.26936712  0.3906775 -0.5150656  0.2997252  -0.03439296
caramel                 0.21311310 -0.1223551  0.3339600 -0.2695850   0.22193335
peanutyalmondy         -0.01764631 -0.2055566  0.2604196 -0.2061093   0.08788927
nougat                 -0.08974359 -0.1386750  0.5229764 -0.3103388   0.12308135
crispedricewafer        1.00000000 -0.1386750  0.4237509 -0.2246934   0.06994969
                  pricepercent winpercent
chocolate            0.5046754  0.6365167
fruity              -0.4309685 -0.3809381
caramel              0.2543271  0.2134163
peanutyalmondy       0.3091532  0.4061922
nougat               0.1531964  0.1993753
crispedricewafer     0.3282654  0.3246797
```

```
corrplot(cij)
```



Q22. Examining this plot what two variables are anti-correlated (i.e. have minus values)?

Fruity and chocolate are the most anti-correlated Pluribus and bar are also anti-correlated

Q23. Similarly, what two variables are most positively correlated?

Winpercent and chocolate are the most positively correlated bar is also highly positively correlated with chocolate candy

```
cij
```

	chocolate	fruity	caramel	peanutyalmondy	nougat
chocolate	1.0000000	-0.74172106	0.24987535	0.37782357	0.25489183
fruity	-0.7417211	1.00000000	-0.33548538	-0.39928014	-0.26936712
caramel	0.2498753	-0.33548538	1.00000000	0.05935614	0.32849280
peanutyalmondy	0.3778236	-0.39928014	0.05935614	1.00000000	0.21311310
nougat	0.2548918	-0.26936712	0.32849280	0.21311310	1.00000000
crispedricewafer	0.3412098	-0.26936712	0.21311310	-0.01764631	-0.08974359
hard	-0.3441769	0.39067750	-0.12235513	-0.20555661	-0.13867505
bar	0.5974211	-0.51506558	0.33396002	0.26041960	0.52297636
pluribus	-0.3396752	0.29972522	-0.26958501	-0.20610932	-0.31033884
sugarpercent	0.1041691	-0.03439296	0.22193335	0.08788927	0.12308135

```
pricepercent      0.5046754 -0.43096853  0.25432709     0.30915323  0.15319643
winpercent        0.6365167 -0.38093814  0.21341630     0.40619220  0.19937530
                crispedricewafer        hard         bar    pluribus
chocolate             0.34120978 -0.34417691  0.59742114 -0.33967519
fruity               -0.26936712  0.39067750 -0.51506558  0.29972522
caramel               0.21311310 -0.12235513  0.33396002 -0.26958501
peanutyalmondy       -0.01764631 -0.20555661  0.26041960 -0.20610932
nougat               -0.08974359 -0.13867505  0.52297636 -0.31033884
crispedricewafer      1.00000000 -0.13867505  0.42375093 -0.22469338
hard                 -0.13867505  1.00000000 -0.26516504  0.01453172
bar                   0.42375093 -0.26516504  1.00000000 -0.59340892
pluribus             -0.22469338  0.01453172 -0.59340892  1.00000000
sugarpercent          0.06994969  0.09180975  0.09998516  0.04552282
pricepercent          0.32826539 -0.24436534  0.51840654 -0.22079363
winpercent            0.32467965 -0.31038158  0.42992933 -0.24744787
                sugarpercent pricepercent winpercent
chocolate          0.10416906    0.5046754  0.6365167
fruity            -0.03439296   -0.4309685 -0.3809381
caramel            0.22193335    0.2543271  0.2134163
peanutyalmondy     0.08788927    0.3091532  0.4061922
nougat             0.12308135    0.1531964  0.1993753
crispedricewafer   0.06994969    0.3282654  0.3246797
hard               0.09180975   -0.2443653 -0.3103816
bar                0.09998516    0.5184065  0.4299293
pluribus           0.04552282   -0.2207936 -0.2474479
sugarpercent       1.00000000    0.3297064  0.2291507
pricepercent       0.32970639    1.0000000  0.3453254
winpercent         0.22915066    0.3453254  1.0000000
```

Section 6: Principal component analysis (PCA)

we can use prcomp and set scale = True since one of the variables is on a different scale than other variables

```
pca <- prcomp(candy, scale = TRUE)
summary(pca)
```

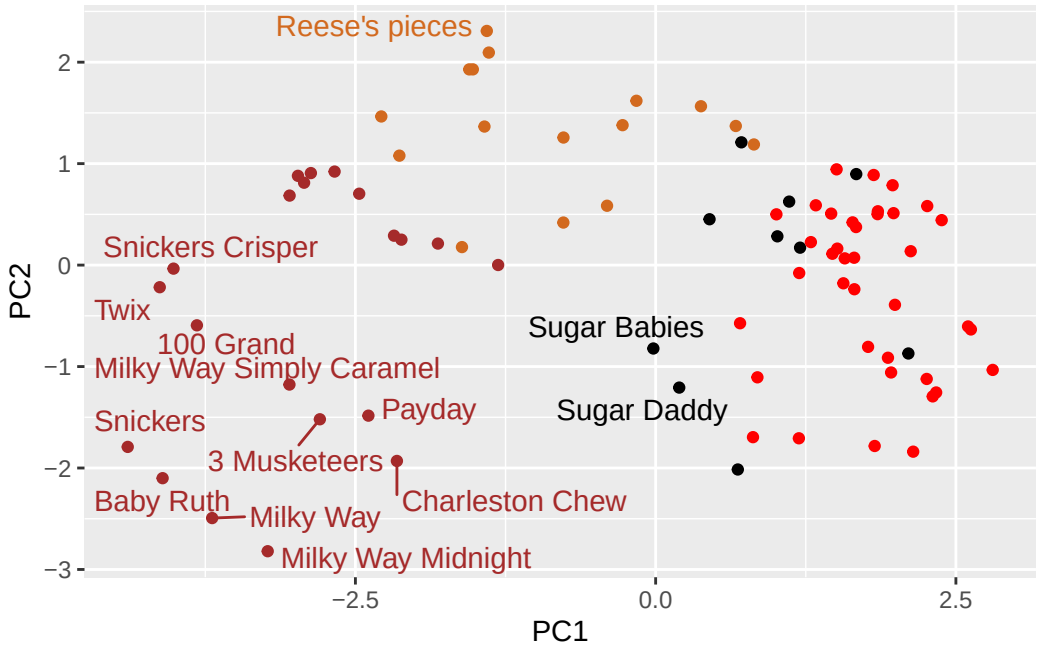
```
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6     PC7
Standard deviation     2.0788 1.1378 1.1092 1.07533 0.9518 0.81923 0.81530
Proportion of Variance 0.3601 0.1079 0.1025 0.09636 0.0755 0.05593 0.05539
```

```
Cumulative Proportion  0.3601 0.4680 0.5705 0.66688 0.7424 0.79830 0.85369
                           PC8     PC9    PC10    PC11    PC12
Standard deviation     0.74530 0.67824 0.62349 0.43974 0.39760
Proportion of Variance 0.04629 0.03833 0.03239 0.01611 0.01317
Cumulative Proportion  0.89998 0.93832 0.97071 0.98683 1.00000
```

Let's make our main results figures, first our score plot (PC plot)

```
ggplot(pca$x) + aes(x=PC1, y = PC2, label=rownames(candy)) + geom_point(color = my_cols) +
  geom_text_repel(col=my_cols, max.overlaps = 5)
```

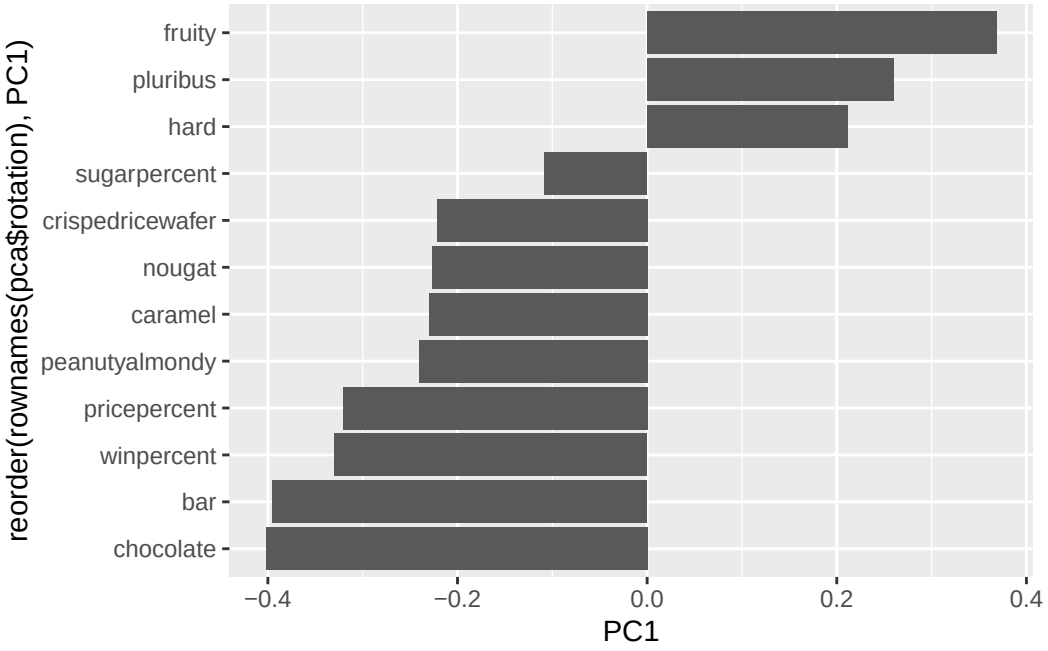
```
Warning: ggrepel: 71 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



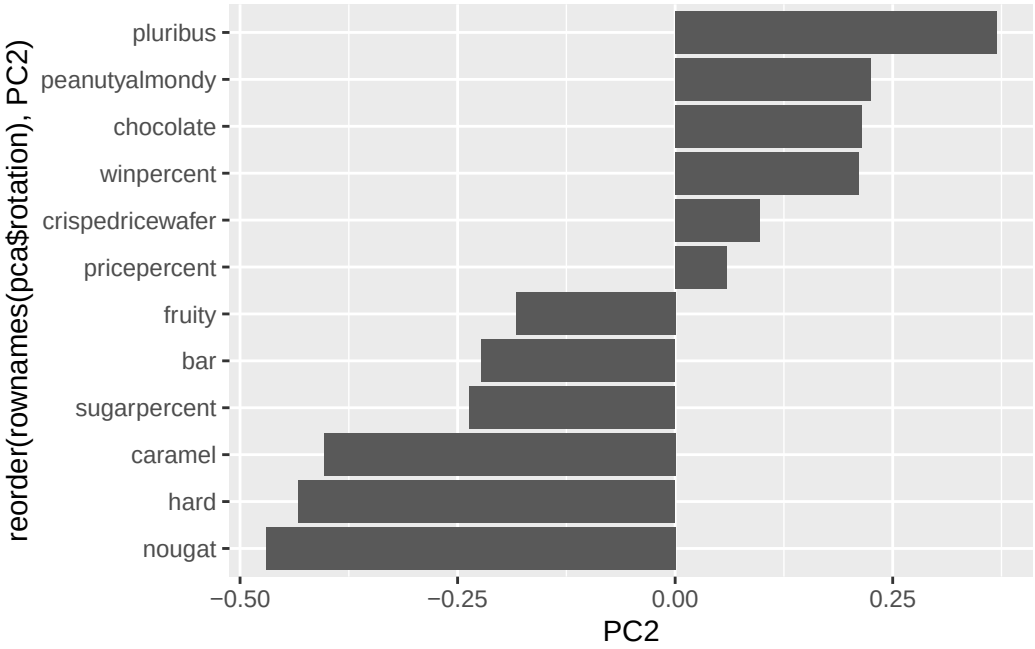
Fruity and chocolate tend to cluster with each other

Let's look at how the original variables contribute to our new PCs. This is often called the variable loadings/contributions

```
ggplot(pca$rotation) +
  aes(x = PC1, y = reorder(rownames(pca$rotation), PC1)) +
  geom_col()
```




```
ggplot(pca$rotation) +
  aes(x = PC2, y = reorder(rownames(pca$rotation), PC2)) +
  geom_col()
```



Q24. What original variables are picked up strongly by PC1 in the positive direction? Do these make sense to you?

Fruity, hard, and pluribus are picked up strongly by PC1 in the positive direction. This makes sense because most pluribus candies are fruity and hard, so it checks out that these variable are picked up together. We also saw that these three variables were positively correlated in the correlation matrix.